Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

ANSWER: Out of 7 categorical variables only year and holiday variables had an effect on the prediction.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

ANSWER: This is important as it reduces the number of variables that needs to be factored in to make the model.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

ANSWER: temp

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

ANSWER:

Linear relationship: As per pair plot it showed that there was a linear regression between X and Y.

Independence: This we factored in through RFE and VIF method

Homoscedasticity: The residuals have constant variance at every level of x.

Normality: The residuals of the model are normally distributed as per plot.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

ANSWER: Year Temperature and Windspeed as their coefficients are high.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

ANSWER:

1. Data Valuation:
    a. Perform EDA to understand variables
    b. Check correlation among the variables
2. Data Preparation
    a. Creating dummy variables.
    b. Divide data to train and test
    c. Scaling
3. Modelling and Evaluation
    a. Create Model
    b. Evaluate the model

2. Explain the Anscombe's quartet in detail. (3 marks)

ANSWER: Set of 4 datasets with similar statistical features but show up differently when represented graphically.

3. What is Pearson's R? (3 marks)
ANSWER: This gives us the strength of relationship between variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
ANSWER: Scaling: This is used to standardize the data to a common unit
Why: It can be used to predict the data in a proper manner and not give an improper result.
Normalized scaling means - X-Min/(Max-Min)
Standardized scaling: X-Mean/SD

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

(3 marks)
ANSWER: This happens when there is a perfect correlation between 2 variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

ANSWER: It is a graphical representation of two data sets coming from populations with a common distribution.